Prosodic and Linguistic Analysis of Semantic Fluency Data: A Window into Speech Production and Cognition

Prosodic and Linguistic Analysis of Semantic Fluency Data: A Window into Speech Production and Cognition

Maria K Wolters[1]*, Najoung Kim*[2]*, Jung-Ho Kim*[2]*, Sarah E. MacPherson*[3]*, Jong C. Park*[2]

[1]School of Informatics, University of Edinburgh
[2]NLP CL Lab, Department of Computer Science, KAIST
[3]Human Cognitive Neuroscience, University of Edinburgh

maria.wolters|sarah.macpherson@ed.ac.uk, park@nlp.kaist.ac.kr (Corresponding Author)

Abstract

Semantic fluency is a commonly used task in psychology that provides data about executive function and semantic memory. Performance on the task is affected by conditions ranging from depression to dementia. The task involves participants naming as many members of a given category (e.g. animals) as possible in sixty seconds. Most of the analyses reported in the literature only rely on word counts and transcribed data, and do not take into account the evidence of utterance planning present in the speech signal. Using data from Korean, we show how prosodic analyses can be combined with computational linguistic analyses of the words produced to provide further insights into the processes involved in producing fluency data. We compare our analyses to an established analysis method for semantic fluency data, manual determination of lexically coherent clusters of words.

Index Terms: verbal fluency, semantic fluency, executive function, prosody, pauses, Korean, word embeddings

1. Introduction

Semantic fluency is widely used in cognitive psychology and neuropsychology. For this task, participants are asked to say as many words that belong to a given category as possible in sixty seconds. The task provides data about executive function, in particular the ability to switch between subcategories, and semantic memory, in particular about structure and richness of the semantic mental lexicon [1, 2, 3, 4]. Semantic fluency is also used in clinical practice, often as part of a standardised test like the Addenbrooke's Cognitive Examination Revised [5] that tracks changes in a person's cognitive status.

In this paper, we focus on a source of information about fluency performance that has been comparatively neglected in the literature, namely prosody. Prosodic measures such as pause duration can be obtained directly from the speech signal without requiring accurate Automatic Speech Recognition (ASR). Here, we focus on pauses and disfluencies due to significant prosodic differences between Korean dialects (cf Sec. 3).

We present pilot results on 40 samples from 20 young native speakers of Korean. For the automatic analysis of semantic fluency data, we used a novel approach based on word embeddings that scales easily to different languages.

We show that prosodic data closely complements word-level analyses. Metrics such as the number of pauses correlate significantly with the number of switches between subcategories as determined both manually and automatically. Pause duration also allows a rough estimate of item response latencies between words. Finally, changes in pause duration over time highlight further potential individual differences in cognitive function.

2. Background

In most of the literature and in clinical practice, performance on semantic fluency tasks is reported as the number of items produced, excluding perseverations and insertions, or the number of words produced in the first, second, third, and fourth 15-second interval. More detailed analyses require audio recordings or writing down each word, and can be quite time-consuming.

Clustering and Switching. These analyses assume that when people produce a sequence of stimuli, they access a particular semantic subcategory, retrieve a series of words from that subcategory, and then switch to a different one when retrieval is exhausted. Perhaps the most well known form of this method was proposed by Troyer and collaborators [1, 6]. The measures can differentiate between older and younger adults, and people with different neurodegenerative disorders [7].

Response Times. These analyses focus on the time taken to search for and access the next lexical item without making assumptions about underlying semantic subcategories. The time t_n between the end of word w_{n-1} and the beginning of word w_n is taken to be the response time. Hesitations, paralinguistic vocalisations such as laughter, and verbal comments are usually regarded as part of response times. Following [8, 2], the response times for one speaker are summarised in a linear model of the form given in Equation 1, where c is a lexical retrieval constant and s models gradual increases in retrieval time. Like clustering and switching, item response times (or latencies) are sensitive to cognitive ageing and cognitive impairment [3].

$$t_n = c + s * n \tag{1}$$

Beyond Manual Annotation. While both clustering and switching and response time analyses are widely used in the psychological literature, they are time consuming to measure and therefore less useful for clinical practice. A variety of automated analysis methods have been proposed that only require a transcription of the words produced, once they have been trained on existing language data. For example, Latent Semantic Analysis [9] has been used to extract clustering information, and lexical similarity measures based on WordNet [10] have been proposed as an alternative to manual scoring.

However, automating transcription of semantic fluency sequences by using ASR is difficult. Pakhomov et al. [11] found that word error rates are high, and that speaker adaptation may be required.

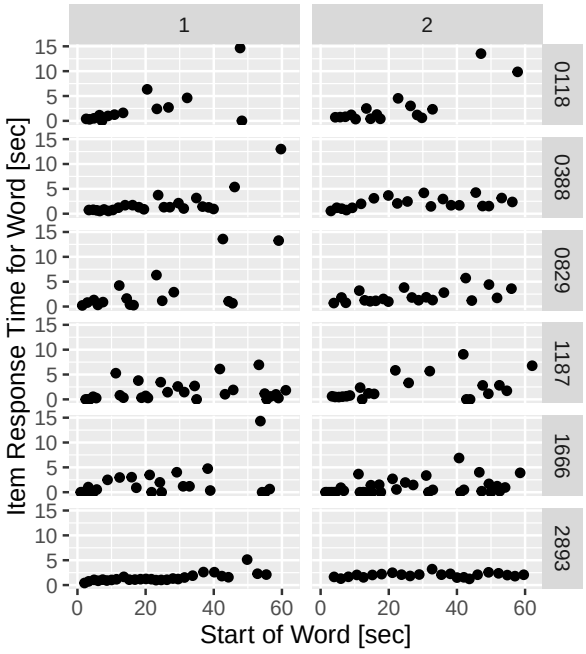

Figure 1: *Sample Item Response Latency Plots for Trial 1 (left) and Trial 2 (right). The item response time for a word is the time between the start of the current word and the end of the previous one. Numbers in the right corner are participant IDs.*

3. Incorporating Prosodic Information

In this paper, we look at three types of prosodic information: articulation rate, disfluencies, and pause patterns.

Articulation rate can be affected by ageing [12] and mental and neurological disorders [13, 14]. For Korean, with its predominantly (C)V(C) syllable structure, we use both the number of words and the number of syllables per second.

Disfluencies are indicators of speech planning processes [15] that are not specifically included in existing analyses. In our Korean material, most disfluencies are filled pauses that are realised as long central nasal vowels.

Pause patterns were chosen for two reasons. First, they approximate lexical access times, where lexical access is not immediate, and second, they are an indicator of phrase boundaries. Pauses are defined as silences longer than 200ms in the speech signal that are not due to obstruent closures.

While Standard Seoul Korean has lost Middle Korean lexical tones, dialects such as the Kyungsang group spoken in the South East [16] have preserved them, and their realisation varies by age group [17]. This makes traditional phrase boundary analyses too complex for the scope of this paper.

4. Data

The dataset used in this paper consists of 40 semantic fluency test sequences collected from 20 university students who were native speakers of Korean for an earlier study of computational models of verbal fluency sequences [18]. Each student was asked to perform the task twice with at least two days of interval period between the experiments. Data from day 1 corresponds to Trial 1, data from day 2 to Trial 2. This allows us to assess practice effects. Data from trials where the participant failed to follow instructions (e.g., producing non-animals) were discarded. All trials were audio recorded and ranscribed by a native Korean speaker. For each participant, we collected demographic information (age and gender) and determined dialect. Since the speech data were only collected to develop automated analysis techniques, no other cognitive tests were performed.

The original number of speakers recorded was 25. Two did not produce usable data, and three did not return for the second recording session, which leaves us with 20 participants who provided two sets of data each, for a total of 40 recordings.

The mean age was 22 (SD: 2; range: 18–27). 6 (30%) participants were female, 14 male. There were no significant differences in age between genders ($p < 0.42$, Fisher's Exact Test). 16 (80%) had a more or less strong standard Seoul accent, which is not tonal, while 4 were speakers of a potentially tonal Kyungsang dialect.

5. Method

5.1. Manual Word-Level Analysis

Clusters were manually annotated by two native speakers of Korean (NJK, JHK) following the established manual scoring method described in [1]. Cohen's κ was 0.67 unweighted (95% CI=[0.62–0.72]) and 0.91 weighted (95% CI=[0.89–0.93]). Since NJK was mainly responsible for developing the automatic analyses, we used JHK's annotations as the basis for the following comparisons.

5.2. Automatic Word-Level Analysis

We propose exploiting word embeddings to automate the process of analyzing and scoring semantic fluency sequences. The initial goal here is to replicate the results obtained from manual scoring of switches between clusters. We used word embeddings instead of the WordNet-derived measures proposed by Pakhomov et al. [10] because these do not require a carefully curated linguistic resource.

We vectorized each word using the popular word2vec model [19] trained on Korean Wikipedia dump from 26 December 2015. We then calculated cluster boundaries using two different approaches. The first approach uses cosine similarity (Eq. 2) between two word vectors $w_{n-1}^{\rightarrow}$ and $\vec{w_n}$ to identify switches and the second approach derives switches from boundaries between clusters that have been determined using vector quantization (k-means). We chose cosine similarity because it is the most popular measure of similarity in vector space models of semantics [20] and vector quantization because the algorithm is designed to identify clusters, which intuitively aligns with our task to automatically find clusters and switches.

$$\cos\theta = \frac{w_{n-1}^{\rightarrow} \cdot \vec{w_n}}{\|w_{n-1}^{\rightarrow}\| \, \|\vec{w_n}\|} \tag{2}$$

We designed the following three cosine similarity-based markers of a switch: 1) fixed threshold value, 2) sharp changes in similarity, and 3) low inter-group similarity. The *fixed threshold value* marks the point where the similarity between two neighboring words fall below a certain fixed value as a switch boundary. The fixed values used here were the median and the 25^{th} percentile of all adjacent cosine similarity values rounded to the nearest second decimal place (0.30 and 0.22 respectively).

The *sharp change* keeps track of the change in cosine similarity between two subsequent words and marks switches where the figure deviates sharply from the average similarity change.

We considered changes that are more than twice as large as the prior average change as sharp.

The *inter-group measure* calculates the average similarity between all possible pairs within a certain sub-sequence, and marks switches where the inter-group similarity is low. We chose 0.30 and 0.36 because this was the median and 75^{th} percentile inter-group similarity (rounded to the nearest second decimal place) of $2 \leq k \leq 5$ animals selected at random, with 1000 samples from each k.

We applied k-means vector quantization to each vectorized word sequence to find clusters of words. The algorithm partitions the set of vectorized words into k clusters, where k is explicitly specified. In our experiment, we tested $2 \leq k \leq 5$ for all 40 vectorized CFT sequences using WEKA [21]. Switching boundaries were marked according to the clusters identified by the algorithm.

5.3. Acoustic and Prosodic Analysis

Words, pauses, disfluencies, laughter, experimenter comments, and participant comments were annotated using Praat 5.4.08 [22]. Pauses were defined as stretches of signal without any vocalisations with a minimum length of 200ms. For words with initial voiceless obstruents, the start of the word was marked at the release of the obstruent closure.

The item response latency of a word w_n was calculated as the time in seconds from the end of the previous word w_{n-1} to the start of w_n. This included pauses, disfluencies, comments, and any other non-word sounds. The constant and slope for each speaker was estimated using a linear mixed model [23] with trial and participant ID as random effects.

5.4. Statistical Analysis

Differences between Trial 1 and Trial 2 were tested for significance using the Asymptotic Wilcoxon-Mann-Whitney Test, R package `coin` [24], and Spearman's ρ was used to compute correlations.

6. Results

6.1. Word-Level Analysis

The participants produced a mean of 23 words per trial (SD: 4.8, range: 14–36). The average number of switches was 11 (SD: 2.9, range: 6–18). There was no significant difference in both measures between Trial 1 and Trial 2 ($p < 0.16$ for word count, $p < 0.47$ for switches).

Table 1 summarises the correlation between the automatic measures and the number of switches determined manually. For both inter-group similarity and the fixed threshold method, we only report the results for the parameter setting that yielded the highest correlation. For both methods, this setting is 0.3, the value derived from the median.

The best correlation is obtained for the fixed threshold. All other methods tend to underestimate the number of switches present by 2–5 on average, whereas the average difference between the number of switches as defined by the median threshold of cosine similarity and the number of switches found by JHK is -0.25 (SD: 3.1). Results from the three best-performing measures do not vary significantly by trial (fixed threshold: $p < 0.29$, vector quantization: $p < 0.96$, inter-group similarity: $p < 0.76$).

Table 1: *Correlation between manually determined switches and automatically determined switches (Spearman's ρ). *: p <0.05, **:p <0.01, ***:p <0.001*

(a) Cosine Similarity	
fixed threshold***	0.53
abrupt change*	0.40
inter-group similarity**	0.43

(b) k-means	
k=2	0.19
k=3*	0.39
k=4*	0.33
k=5	0.26

Table 2: *Descriptive Statistics for Item Response Time Analysis.*

		M	SD	Range
Values	First	0.55	0.53	0.00–2.80
	Median	0.74	0.46	0.00–2.67
	3^{rd} Quartile	2.49	0.94	0.99–6.22
	Max	8.64	3.65	3.2–15.8
Model	Constant	1.24	0.25	0.59–1.47
	Slope	0.00	0.2	-0.02–0.08

6.2. Acoustic and Prosodic Analysis

Table 2 summarises item response time statistics. The lag between the first and the second word and the median lag between words are both relatively short compared to the third quartile. This indicates that for most speakers, we see mostly brief response times, interspersed with a few long ones.

The latency between first and second word and the median latency are both significantly lower in Trial 2 (first: $p < 0.003$, mean Δ -0.42, 95% confidence interval [-0.69,-0.19]; median: $p < 0.05$, mean Δ -0.24, 95% CI [-0.46,-0.04]).

Looking at the linear mixed model, we find that the intercept (the constant in Eq. 1) shows substantial variation, while the slope is nearly flat. Fig. 1 plots the time course of response latencies for six participants for Trial 1 and Trial 2. The basic pattern consists of latencies that vary randomly across a constant, as the model suggests. In addition, however, many participants show several clear outliers. These are responsible for the substantial variation in third quartiles and maximum latency values shown in Table 2.

Of the descriptive latency statistics, only the third quartile correlates with the manual number of switches ($\rho = -0.33$, $p < 0.04$), all others do not (median: $\rho = 0.10$, first latency $\rho = 0.12$, maximum latency $\rho = -0.25$). This suggests that participants who deviate more from the constant pattern of latencies may find it harder to switch between subcategories. Both the third quartile ($\rho = -0.50$, $p < 0.005$) and the maximum latency ($\rho = -0.41$, $p < 0.01$) also correlate with the number of switches determined using the fixed threshold method.

Table 3 summarises the results of the prosodic analysis. Of the prosodic measures, only the number of disfluencies is affected by trial; later trials are less disfluent. Both the number of pauses and the articulation rate in words per second correlate positively with the number of manually determined switches. The correlations between the number of switches determined automatically using a fixed threshold for cosine similarity and the various prosodic measures listed in Table 3 are even higher. This suggests that the number of words produced may be a mediating variable here.

Table 3: *Prosodic Analysis. Descriptive Statistics, differences between trials (Asymptotic Wilcoxon-Mann-Whitney test) and correlation with number of switches (Spearman's ρ). Automatic switches determined by fixed threshold method. *: $p < 0.05$, **:$p < 0.01$, ***:$p < 0.001$*

Measure		M	SD	Range	Δ_{trial} (95% CI)	ρ_{switch}	
						Man.	Auto
No. Disfluencies		2.45	3.15	0–12	1* [0,4]	-0.11	-0.07
No. Pauses		23.45	5.53	11–33	-1 [-4,3]	0.45**	0.51***
Pauses	Median	1.33	0.49	0.66–3.11	-0.02 [-0.28,0.18]	-0.14	-0.32*
in sec	Max	7.12	3.19	2.59–15.78	1.15 [-0.56,3.16]	-0.10	-0.40**
Rate	Word	1.75	0.26	1.24–2.21	-0.02 [-0.20,0.15]	0.45**	0.60***
in 1/sec	Syllable	4	0.55	3.00–5.32	-0.14 [-0.52,0.22]	0.27	0.31*

7. Discussion

7.1. Information in the Speech Signal

Word-level lexical analysis and acoustic / prosodic analysis provide complementary sources of information about the internal structure of verbal fluency data. Pause durations contain information about lexical access times, are less sensitive to practice effects than item response times, and correlate with the number of switches detected manually. This fits well with the literature. While cluster and switching based models and item response time models have somewhat different theoretical underpinnings, in practice, their results are fairly similar [3].

Our young participants vary substantially in their response latency patterns. For some participants, item response times show additional peaks over the course of the utterance that are typically followed by several shorter response latencies. Abwender et al. [25] suggested an additional type of switch that might account for these patterns, *hard switches*, where participants produce a sequence of one-word clusters. Initial inspection of the data suggests that hard switches cannot explain all of the variation seen. For example, participants like 2893, Trial 2, (Fig. 1) show an almost level latency pattern, even though most of the switches are hard switches. Thus, it remains to be seen whether these differences in response patterns are mainly due to executive function [2] (i.e., difficulties with switching to new lexical categories) or to slower lexical access times for some parts of semantic memory.

The speech data would also benefit from prosodic annotation using KToBI system [26]. This would give us information about the type of boundary tone that corresponds to longer response latencies and their communicative function [27]. Such an analysis needs to be sensitive to dialect.

Looking at speaking and articulation rates, we find that participants produce far more words in the first quarter of the utterance than in each of the remaining quarters. For some participants, this decline is sharper than for others. Such a pattern is normal [28, 29, 30]. One potential reason is that participants first retrieve clusters associated with relatively frequent words, and then require longer to search for less frequent words [28, 30]. As Raboutet et al. [30] found, using traditional normative corpus data to estimate frequencies is problematic if the underlying corpus is not well-balanced.

Finally, it is worth noting that some of the non-speech events we found in the data, and that have also been reported in other studies [3] (e.g., laughter, comments) may have occurred because the data was collected face-to-face. Before setting up a mechanism for administering the task remotely, as in [31], sample data needs to be collected in both communication situations.

7.2. Towards Combined Automated Analysis

Both word-level analysis and prosodic analysis are comparatively language-independent and easy to automate. Our *linguistic* analysis approach relies on deriving similarities from word embeddings. Unlike some researchers [32, 33], who have extensively used clustering algorithms, we are not looking to characterise the organisation of semantic memory. We chose to use a context prediction-based model of distributional word representation rather than count-based models like Latent Semantic Analysis (LSA), as the former performs better than the latter in various NLP tasks such as measuring semantic relation similarity and syntactic regularities [34, 35].

The acoustic and prosodic analysis is also relatively robust and easy to implement. If recording quality is high and background noise is low, automated pause detection is comparatively straightforward. Information about pausing patterns might also complement ASR analyses. Fully replicating our approach requires detectors for disfluencies and relevant non-speech events such as laughter, which will be useful when semantic fluency is used to track changes in cognitive function due to mental health [28, 36].

8. Conclusion

We have shown that prosodic and lexical-level analyses of semantic fluency data complement each other well. Since our baseline findings on this data set fit well with what we know from the literature about semantic fluency, we are reasonably confident that our findings will hold for larger, more diverse data sets. In future, we plan to record fluency data from a larger sample that varies along dimensions known to affect executive function and fluency performance, such as age and mental health.

We have also highlighted patterns in item response latencies that require further investigation in a study that uses detailed psychological testing in combination with full prosodic and phonetic annotations to identify potential reasons and investigate relevant aspects of intonation and speech rhythm.

Finally, our acoustic analysis relies completely on manual annotation. It would be interesting to see whether the findings can be replicated with robust detection algorithms for pauses and paralinguistic vocalisations.

9. Acknowledgements

This work was supported by the National Research Foundation of Korea grant funded by the Korea government (MSIP) (No. 2010-0028631) and the Challenge Investment Fund (No. G790CH) of the University of Edinburgh. We thank Akira Utsugi for useful comments.

10. References

[1] A. K. Troyer, M. Moscovitch, and G. Winocur, "Clustering and switching as two components of verbal fluency: evidence from younger and older healthy adults." *Neuropsychology*, vol. 11, no. 1, pp. 138–46, 1997.

[2] U. Mayr and R. Kliegl, "Complex semantic processing in old age: does it stay or does it go?" *Psychology and Aging*, vol. 15, no. 1, pp. 29–43, 2000.

[3] J. McDowd, L. Hoffman, E. Rozek, K. E. Lyons, R. Pahwa, J. Burns, and S. Kemper, "Understanding verbal fluency in healthy aging, Alzheimer's disease, and Parkinson's disease." *Neuropsychology*, vol. 25, no. 2, pp. 210–25, 2011.

[4] Z. Shao, E. Janse, K. Visser, and A. S. Meyer, "What do verbal fluency tasks measure? Predictors of verbal fluency performance in older adults," *Frontiers in Psychology*, vol. 5, no. July, pp. 1–10, 2014.

[5] E. Mioshi, K. Dawson, J. Mitchell, R. Arnold, and J. R. Hodges, "The Addenbrooke's Cognitive Examination Revised (ACE-R): a brief cognitive test battery for dementia screening." *Int J Geriatr Psychiatry*, vol. 21, no. 11, pp. 1078–1085, 2006.

[6] A. K. Troyer, "Normative data for clustering and switching on verbal fluency tasks." *Journal of Clinical and Experimental Neuropsychology*, vol. 22, no. 3, pp. 370–8, 2000.

[7] A. K. Troyer, M. Moscovitch, G. Winocur, L. Leach, and M. Freedman, "Clustering and switching on verbal fluency tests in Alzheimer's and Parkinson's disease." *Journal of the International Neuropsychological Society*, vol. 4, no. 2, pp. 137–43, 1998.

[8] D. Rohrer, J. T. Wixted, D. P. Salmon, and N. Butters, "Retrieval from semantic memory and its implications for Alzheimer's disease." *Journal of Experimental Psychology: Learning, Memory, and Cognition*, vol. 21, no. 5, pp. 1127–39, 1995.

[9] K. K. Nicodemus, B. Elvevåg, P. W. Foltz, M. Rosenstein, C. Diaz-Asper, and D. R. Weinberger, "Category fluency, latent semantic analysis and schizophrenia: a candidate gene approach." *Cortex*, vol. 55, pp. 182–91, 2014.

[10] S. V. S. Pakhomov, L. S. Hemmy, and K. O. Lim, "Automated semantic indices related to cognitive function and rate of cognitive decline." *Neuropsychologia*, vol. 50, no. 9, pp. 2165–75, 2012.

[11] S. V. Pakhomov, S. E. Marino, S. Banks, and C. Bernick, "Using Automatic Speech Recognition to Assess Spoken Responses to Cognitive Tests of Semantic Verbal Fluency," *Speech Communication*, vol. 75, pp. 14–26, 2015.

[12] B. L. Smith, J. Wasowicz, and J. Preston, "Temporal characteristics of the speech of normal elderly adults," *Journal of Speech and Hearing Research*, vol. 30, no. 4, pp. 522–529, 1987.

[13] J. Duffy, *Motor Speech Disorders*. Mosby, 1995.

[14] H. Ellgring and K. R. Scherer, "Vocal indicators of mood change in depression," *Journal of Nonverbal Behavior*, vol. 20, no. 2, pp. 83–110, 1996.

[15] H. Bortfeld, S. D. Leon, J. E. Bloom, M. F. Schober, and S. E. Brennan, "Disfluency rates in conversation: Effects of age, relationship, topic, role, and gender," *Language and Speech*, vol. 44, pp. 123–147, 2001.

[16] M. Kenstowicz and C. Park, "Laryngeal features and tone in kyungsang korean: A phonetic study," *Studies in Phonetics, Phonology, and Morphology*, vol. 12, no. 2, pp. 247–264, 2006.

[17] A. Utsugi, "Merger-in-Progress of Tonal Classes in Masan/Changwon Korean." *Language Research*, vol. 45, no. 1, pp. 23–42, 2009.

[18] Y. J. Lee, H. J. Lee, M. Wolters, and J. C. Park, "Analyzing the patterns of switching and clustering on CFT data using hidden markov model," *HCI 2012*, pp. 181–184, 2012.

[19] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in neural information processing systems*, 2013, pp. 3111–3119.

[20] P. D. Turney and P. Pantel, "From frequency to meaning: Vector space models of semantics," *Journal of Artificial Intelligence Research*, vol. 37, no. 1, pp. 141–188, 2010.

[21] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The weka data mining software: an update," *ACM SIGKDD explorations newsletter*, vol. 11, no. 1, pp. 10–18, 2009.

[22] P. Boersma and D. Weenink, *Praat*, 2015.

[23] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting linear mixed-effects models using lme4," *Journal of Statistical Software*, vol. 67, no. 1, pp. 1–48, 2015.

[24] T. Hothorn, K. Hornik, M. A. van de Wiel, and A. Zeileis, "Implementing a class of permutation tests: The coin package," *Journal of Statistical Software*, vol. 28, no. 8, pp. 1–23, 2008.

[25] D. A. Abwender, J. G. Swan, J. T. Bowerman, and S. W. Connolly, "Qualitative analysis of verbal fluency output: review and comparison of several scoring methods." *Assessment*, vol. 8, no. 3, pp. 323–38, 2001.

[26] S.-A. Jun, "K-Tobi Labelling Conventions," *UCLA Working Papers in Phonetics*, vol. 99, pp. 149–173, 2000.

[27] H.-S. Jeon, "The Perception of Korean Boundary Tones by First and Second Language Speakers." in *Speech Prosody 7*, 2014.

[28] S. F. Crowe, "The Performance of Schizophrenic and Depressed Subjects on Tests of Fluency: Support for a Compromise in Dorsolateral Prefrontal Functioning," *Australian Psychologist*, vol. 31, no. 3, pp. 204–209, 1996.

[29] ——, "Decrease in performance on the verbal fluency test as a function of time: evaluation in a young healthy sample." *Journal of clinical and experimental neuropsychology*, vol. 20, no. 3, pp. 391–401, 1998.

[30] C. Raboutet, H. Sauzeon, M.-M. Corsini, J. Rodrigues, S. Langevin, and B. N'kaoua, "Performance on a semantic verbal fluency task across time: dissociation between clustering, switching, and categorical exploitation processes." *Journal of Clinical and Experimental Neuropsychology*, vol. 32, no. February 2015, pp. 268–280, 2010.

[31] M. K. Wolters, L. Ferrini, E. Farrow, A. Szentagotai Tatar, and C. D. Burton, "Tracking depressed mood using speech pause patterns," in *International Congress of Phonetic Sciences*, 2015.

[32] T. Prescott, L. D. Newton, N. U. Mir, P. W. R. Woodruff, and R. Parks, "A new dissimilarity measure for finding semantic structure in category fluency data with implications for understanding memory organization in schizophrenia," *Neuropsychology*, vol. 20, pp. 685–699, 2006.

[33] C. Sumiyoshi, A. Ertugrul, A. E. Anil Yagcioglu, and T. Sumiyoshi, "Semantic memory deficits based on category fluency performance in schizophrenia: similar impairment patterns of semantic organization across Turkish and Japanese patients." *Psychiatry research*, vol. 167, no. 1-2, pp. 47–57, 2009.

[34] M. Baroni, G. Dinu, and G. Kruszewski, "Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors." in *ACL*, 2014, pp. 238–247.

[35] T. Mikolov, W.-T. Yih, and G. Zweig, "Linguistic regularities in continuous space word representations." in *HLT-NAACL*, 2013, pp. 746–751.

[36] J. Henry and J. R. Crawford, "A meta-analytic review of verbal fluency deficits in depression." *Journal of Clinical and Experimental Neuropsychology*, vol. 27, no. 1, pp. 78–101, 2005.